CARRIER-PERTURBATION RESPONSES: AN ALTERNATIVE BASIS FOR GENERALIZABLE AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing generated image detectors typically infer image authenticity from features extracted from the image itself. However, such features may vary substantially across scenes, limiting generalization from unseen domains. Our preliminary observations show that, under the same certain controlled perturbations, real and generated images exhibit different feature-change distributions in the representation space of a frozen visual foundation model, while responses from same-class images remain relatively consistent across scenes. This motivates us to investigate perturbation-induced feature changes as an alternative basis for authenticity assessment. To this end, we propose **Carrier-Perturbation Responses (CPR)**, which treats each image as a carrier and characterizes its response to a controlled perturbation through the feature change between the original and perturbed images. Since some perturbations may produce responses dominated by the perturbation rule itself or responses too weak for discrimination, we further develop an ANOVA-inspired selection criterion that favors perturbations with strong carrier–perturbation interactions, limited rule-induced effects, and sufficient response magnitude. Together, these designs shift the detection basis from scene-dependent appearance features toward perturbation-induced response patterns, providing complementary evidence for authenticity assessment across unseen scenes. Across four in-the-wild datasets, CPR improves the average real-image accuracy by 3.3 percentage points to 96.7% and the average overall accuracy by 2.1 percentage points to 95.4%. Our code is available at https://anonymous.4open.science/r/CPR-F211.

1 INTRODUCTION

The rapid advancement of image generation models (Karras et al., 2017; Team et al., 2023; Goodfellow et al., 2014; Karras et al., 2019) has made it increasingly difficult to determine whether visual content originates from the real world or generation models. Reliable generated image detection is therefore essential for content authenticity verification (Li et al., 2025b).

Representative detectors generally follow two detection paradigms. The first employs conventional convolutional neural networks (CNNs) (Simonyan & Zisserman, 2014; He et al., 2016) to identify generation artifacts in pixels, textures, or frequency components (Afchar et al., 2018; Liu et al., 2020; Frank et al., 2020). Under this paradigm, real images are typically identified indirectly through the absence of detected generation artifacts. When learned artifact cues fail to transfer to out-of-domain generated images, these images may be misclassified as real. For example, NPR (Tan et al., 2024) achieves a real-image accuracy of 99.9% on Chameleon (Yan et al., 2025), but its generated-image accuracy is only 4.6%.

The second paradigm exploits knowledge encoded in frozen pretrained vision models (Radford et al., 2021) to assess whether an image is consistent with the models' representations of the real world. However, images from unseen scenes may deviate from the representations observed during detector training, and such scene-induced deviations may cause real images to be misclassified as generated. For example, when transferring from the in-domain Stable Diffusion v1.4 (Zhu et al., 2023) test set

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

to Chameleon, the real-image accuracy of Effort (Yan et al., 2024) drops sharply from 88.3% to 39.4%.

The respective bases for authenticity assessment in these two paradigms may lead to predictions biased toward different classes on out-of-domain data. Despite their different detection strategies, both paradigms typically infer authenticity from features extracted from the input image itself. Recent studies show that even a linear classifier trained on frozen features from modern visual foundation models (Bolya et al., 2026; Chuang et al., 2026; Siméoni et al., 2025) can effectively distinguish real and generated images (Zhou et al., 2026). Nevertheless, this approach still relies on features extracted from the image itself, which may vary substantially across scenes and thus limit generalization to unseen domains. This motivates us to explore an alternative basis for authenticity assessment that can provide complementary evidence across unseen scenes.

Our preliminary observations show that, under the same certain controlled perturbations, real and generated images exhibit different feature-change distributions in the representation space of a frozen visual foundation model. Our visualizations further show more compact within-class clustering across scenes for these feature changes than for the original image features. These observations motivate us to investigate whether perturbation-induced feature changes can provide an alternative basis for detection.

To this end, we propose **Carrier-Perturbation Responses (CPR)**, which treats each image as a carrier and characterizes the visual model's response to a controlled perturbation through the feature change between the original and perturbed images. Specifically, a frozen visual foundation model extracts features from both images, and their difference constitutes the response on the given carrier. By applying perturbations under the same rule across images, CPR characterizes how the model responds to a shared probing operation on different carriers. These responses serve as an alternative basis for authenticity assessment, shifting the focus from scene-dependent appearance features toward perturbation-induced response patterns.

Not all perturbation categories yield sufficiently discriminative CPRs. Some perturbations may induce responses dominated by the perturbation rules themselves, while others may produce responses too weak in magnitude for discrimination. We therefore develop an ANOVA-inspired (Montgomery, 2010; Fisher, 1930) selection criterion that estimates the relative contributions of the carrier effect, perturbation effect, and carrier–perturbation interaction effect to variation in CPRs. These contributions are considered together with response magnitude to favor perturbation categories with strong carrier–perturbation interactions, limited rule-induced effects, and sufficient response magnitude. The perturbation categories are selected independently from multiple perspectives, and their CPR-based predictions are then aggregated.

We evaluate CPR on standard generated image detection benchmarks and four in-the-wild datasets. CPR improves the average real-image accuracy on the in-the-wild datasets from 93.4% to 96.7% and the average overall accuracy from 93.3% to 95.4%. These results demonstrate the effectiveness of CPR-based detection in the evaluated cross-domain settings.

Our main contributions are summarized as follows:

- We propose Carrier-Perturbation Responses (CPR), which uses feature changes between perturbed and original images to represent visual foundation model's responses to perturbations on image carriers, providing an alternative basis for authenticity assessment.
- We develop an ANOVA-inspired perturbation selection criterion that considers the relative contributions of carrier, perturbation, and carrier–perturbation interaction effects together with response magnitude, favoring perturbations with strong interactions, limited rule-induced effects, and sufficient response magnitude.
- We evaluate CPR on standard generated image detection benchmarks and four in-the-wild datasets, demonstrating its effectiveness in cross-domain detection.

2 Related Work

CNN-based artifact detection. Early generated image detectors use CNNs to identify generation artifacts in pixels, textures, or frequency statistics (Afchar et al., 2018; Frank et al., 2020). Gram-Net

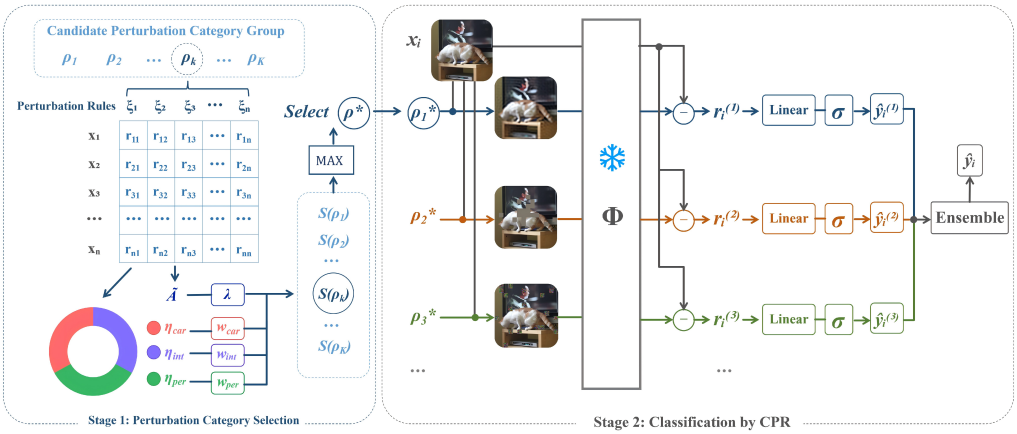


Figure 1: Overview of the proposed Carrier–Perturbation Response (CPR) framework. Stage 1 selects a discriminative perturbation category for each candidate group via the ANOVA-inspired variance decomposition and the weighted selection score $S(\rho_k)$. Stage 2 constructs the CPR $\mathbf{r}_i = \Phi(x_i + \delta_i) - \Phi(x_i)$ for each selected perturbation and trains a linear classification head on the frozen encoder features. The head outputs are averaged to produce the final ensemble prediction.

(Liu et al., 2020) models global texture statistics, while NPR exploits neighboring-pixel relationships introduced by neural upsampling. CNNSpot (Wang et al., 2020) similarly learns transferable artifacts from CNN-generated images. These methods generally identify real images indirectly through the absence of learned generation artifacts. When such artifacts fail to transfer to unseen generators, generated images may consequently be misclassified as real, leading to strongly one-sided predictions.

Detection with pretrained vision models. A separate line of work leverages representations learned by pretrained vision models (Baraldi et al., 2024; Liu et al., 2024). UnivFD (Ojha et al., 2023) trains a lightweight classifier on frozen CLIP (Radford et al., 2021) features, whereas Effort (Yan et al., 2024) decomposes pretrained features into orthogonal subspaces to emphasize generation-related information. More recently, frozen features from modern visual foundation models (Bolya et al., 2026; Chuang et al., 2026), particularly DINOv3 (Siméoni et al., 2025), have shown strong separability using only a linear classifier (Zhou et al., 2026). Nevertheless, these methods infer authenticity from features extracted from the image itself, which may vary across scenes and thus limit generalization to unseen domains.

Reconstruction-based detection. Another line of work uses reconstruction errors as a basis for authenticity assessment. DIRE (Wang et al., 2023) reconstructs an input image with a diffusion model (Ho et al., 2020; Dhariwal & Nichol, 2021) and uses the reconstruction residual for detection. AEROBLADE (Ricker et al., 2024) instead uses autoencoder reconstruction error to detect latent-diffusion-generated images. These methods derive authenticity evidence from how well a generative model reconstructs an input image, under this criterion, images that are difficult to reconstruct may be more likely to be classified as real, making detection vulnerable to scene shifts. In contrast, CPR characterizes a frozen visual foundation model's responses to perturbations through feature changes.

3 METHOD

We propose **Carrier–Perturbation Responses (CPR)** as an alternative basis for generated image detection. CPR treats each image as a carrier and uses a frozen visual foundation model to characterize its responses to controlled perturbations through feature changes between the original and perturbed images. We select perturbation categories and train lightweight linear classifiers on their corresponding CPRs, aggregating the predictions to assess image authenticity.

3.1 CARRIER–PERTURBATION RESPONSE CONSTRUCTION

Conventional generated image detectors encode an input image $x_i \in \mathcal{X}$ as

$$\mathbf{z}_i = \Phi(x_i), \tag{1}$$

where $\Phi(\cdot)$ denotes a frozen visual foundation model. The authenticity of x_i is then inferred from $\mathbf{z}_i$. Since $\mathbf{z}_i$ is computed directly from the input image, it may vary with the scene, which can limit the generalization of the resulting detector under scene distribution shifts. This motivates us to seek an alternative basis for authenticity assessment.

To explore this possibility, we examined how the feature representation changes when a controlled perturbation is applied to an image. We treat x_i as a carrier and apply a perturbation according to a specified category and rule. Let ρ denote a perturbation category and let ξ denote the corresponding rule parameters. The perturbation instance generated for x_i is written as

$$\delta_i = \rho(x_i; \xi), \tag{2}$$

and the perturbed image is given by

$$\tilde{x}_i = x_i + \delta_i, \tag{3}$$

where x_i serves as the carrier of δ_i.

The frozen visual foundation model produces features for both the original and perturbed images:

$$\mathbf{z}_i = \Phi(x_i), \qquad \tilde{\mathbf{z}}_i = \Phi(\tilde{x}_i) = \Phi(x_i + \delta_i). \tag{4}$$

We define the feature change induced by the perturbation as

$$\mathbf{r}_i = \tilde{\mathbf{z}}_i - \mathbf{z}_i = \Phi(x_i + \delta_i) - \Phi(x_i). \tag{5}$$

We refer to $\mathbf{r}_i$ as the **Carrier–Perturbation Response (CPR)**. It characterizes the change in the visual foundation model's features after the perturbation is applied to the carrier image, and thus describes the model's response to the perturbation under the corresponding carrier condition.

After constructing CPRs, we observe that certain perturbations, such as $\rho_{S,4}$ (Sec. 4.2), yield distinct responses for real and generated images. Table 1 quantifies this observation: compared with the original features, CPR yields a smaller cross-domain centroid distance for same-class images. It also increases the real–fake centroid distance and improves both shift-specific AUROCs. Metric definitions are provided in Appendix E. Together with the qualitative visualization in Fig. 2, compared with the original image representations, CPRs from same-class images across seen and unseen scenes form more compact clusters. These observations motivate us to explore CPR as an alternative basis for authenticity assessment.

3.2 PERTURBATION CATEGORY SELECTION

Let $\mathcal{R} = \{\rho_k\}_{k=1}^{K}$ denote the set of candidate perturbation categories. The CPR generated by a given perturbation category is jointly influenced by the carrier effect, the perturbation effect, and the carrier–perturbation interaction effect, where the perturbation effect represents the influence of the perturbation rule. CPRs generated by different perturbation categories may exhibit different effect compositions. This motivates a variance-based approximation, in which the relative contributions of these effects to the observed variation in CPRs serve as proxy measures for perturbation category selection.

3.2.1 VARIANCE-BASED RESPONSE DECOMPOSITION

To obtain an approximate characterization of the effect composition of CPRs, we use an ANOVA-inspired (Montgomery, 2010; Fisher, 1930) repeated-measures variance decomposition. For each candidate perturbation category $\rho_k \in \mathcal{R}$, the observed CPR variation is decomposed into carrier effects, perturbation effects, and carrier–perturbation interaction effects, where the perturbation effect represents the influence of the perturbation rule.

For each perturbation category ρ_k, we conduct G experiments. Each experiment contains N carrier images and N perturbation rules generated from N random parameter seeds under ρ_k. The j-th column shares the rule ξ_{gj}, and the perturbation is independently generated for each carrier as

$$\delta_{gij}^{(k)} = \rho_k(x_{gi}; \xi_{gj}), \qquad g = 1, \ldots, G, \quad j = 1, \ldots, N. \tag{6}$$

Table 1: Quantitative comparison of original features and CPR. Distances are computed in L2-normalized feature spaces before t-SNE. AUROC subscripts indicate the class shifted to the unseen domain. Seen real/fake and unseen fake images are from GenImage (Zhu et al., 2023). Unseen real images are from COCO (Lin et al., 2014) and OpenImages (Kuznetsova et al., 2020), rather than GenImage's ImageNet real-image source (Deng et al., 2009).

Representation	$D_{\text{real}}^{\text{cross}}$	$D_{\text{fake}}^{\text{cross}}$	D_{class}	$\text{AUROC}_{\text{real}}$	$\text{AUROC}_{\text{fake}}$
Original	0.0532 ± 0.0041	0.0678 ± 0.0037	0.0959 ± 0.0023	0.9420 ± 0.0177	0.9338 ± 0.0089
CPR	0.0449 ± 0.0030	0.0661 ± 0.0033	0.1396 ± 0.0026	0.9675 ± 0.0065	0.9408 ± 0.0063



(a) Original features (b) CPR

Figure 2: t-SNE visualization of original features and CPR.

The corresponding response is

$$\mathbf{r}_{gij}^{(k)} = \Phi\left(x_{gi} + \delta_{gij}^{(k)}\right) - \Phi(x_{gi}). \tag{7}$$

Let $\bar{\mathbf{r}}_{gi\cdot}^{(k)}$, $\bar{\mathbf{r}}_{g\cdot j}^{(k)}$, and $\bar{\mathbf{r}}_{g\cdot\cdot}^{(k)}$ denote the mean response over perturbation rules, carriers, and all observations within the g-th experiment, respectively. The corresponding sums of squares are defined as

$$\begin{aligned} SS_{\text{car}}^{(k)} &= N \sum_{g=1}^{G} \sum_{i=1}^{N} \left\| \bar{\mathbf{r}}_{gi\cdot}^{(k)} - \bar{\mathbf{r}}_{g\cdot\cdot}^{(k)} \right\|_2^2, \qquad SS_{\text{per}}^{(k)} = N \sum_{g=1}^{G} \sum_{j=1}^{N} \left\| \bar{\mathbf{r}}_{g\cdot j}^{(k)} - \bar{\mathbf{r}}_{g\cdot\cdot}^{(k)} \right\|_2^2, \\ SS_{\text{int}}^{(k)} &= \sum_{g=1}^{G} \sum_{i=1}^{N} \sum_{j=1}^{N} \left\| \mathbf{r}_{gij}^{(k)} - \bar{\mathbf{r}}_{gi\cdot}^{(k)} - \bar{\mathbf{r}}_{g\cdot j}^{(k)} + \bar{\mathbf{r}}_{g\cdot\cdot}^{(k)} \right\|_2^2. \end{aligned} \tag{8}$$

Here, the three sums of squares quantify the response variation associated with carrier effects, perturbation effects, and their interaction, respectively. Each sum of squares is obtained by aggregating the squared response deviations corresponding to one of these effects. A larger sum of squares indicates a larger contribution of the corresponding effect to the total response variation. Accordingly, the ratio of an individual sum of squares to the total sum of squares provides a normalized measure of its relative contribution to the observed response variation.

The total response variation is measured by

$$SS_{\text{tot}}^{(k)} = \sum_{g=1}^{G} \sum_{i=1}^{N} \sum_{j=1}^{N} \left\| \mathbf{r}_{gij}^{(k)} - \bar{\mathbf{r}}_{g\cdot\cdot}^{(k)} \right\|_2^2, \tag{9}$$

which satisfies

$$SS_{\text{tot}}^{(k)} = SS_{\text{car}}^{(k)} + SS_{\text{per}}^{(k)} + SS_{\text{int}}^{(k)}. \tag{10}$$

Finally, the normalized contributions are computed as

$$\eta_q^{(k)} = \frac{SS_q^{(k)}}{SS_{\text{tot}}^{(k)} + \epsilon}, \qquad q \in \{\text{car}, \text{per}, \text{int}\}, \tag{11}$$

where ϵ is a small constant for numerical stability.

These normalized contributions are used to approximately measure the relative prominence of the carrier effect, perturbation effect, and carrier–perturbation interaction effect in the CPRs, based on their respective contributions to the observed response variation.

3.2.2 SELECTION CRITERION

To select a perturbation category that produces discriminative CPRs, we consider both the effect components defined in Sec. 3.2.1 and the average CPR strength. For each candidate perturbation category ρ_k, the average CPR strength is computed and normalized as

$$A^{(k)} = \frac{1}{|\mathcal{C}_k|} \sum_{\mathbf{r} \in \mathcal{C}_k} \|\mathbf{r}\|_2, \qquad \widetilde{A}^{(k)} = \frac{A^{(k)}}{\max_{\rho_l \in \mathcal{R}} A^{(l)} + \epsilon}, \tag{12}$$

where $\mathcal{C}_k$ denotes the set of CPRs generated using ρ_k, $\mathcal{R}$ is the set of candidate perturbation categories, and ϵ is a small constant for numerical stability.

The selection score of ρ_k is then defined as

$$S(\rho_k) = w_{\text{int}} \eta_{\text{int}}^{(k)} + w_{\text{car}} \eta_{\text{car}}^{(k)} + w_{\text{per}} \eta_{\text{per}}^{(k)} + \lambda \widetilde{A}^{(k)}. \tag{13}$$

Priority-guided weighting. Based on the expected roles of the different effects, we set the priority relation $w_{\text{int}} > w_{\text{car}} > w_{\text{per}}$. The carrier effect characterizes response variation associated with carrier differences and thus contributes to carrier classification. The interaction effect further captures the joint variation between carriers and perturbation rules, contributing to carrier classification and improving model generalization. Therefore, it receives the highest positive weight. In contrast, the perturbation effect may reflect response variation related to specific perturbation rules but irrelevant to carrier classification, so its weight is assigned a relatively low or negative value. The scoring function also incorporates the normalized average CPR strength to prevent weak-response perturbation categories from being selected solely due to their effect composition.

Finally, the optimal perturbation category is selected by

$$\rho^* = \arg\max_{\rho_k \in \mathcal{R}} S(\rho_k). \tag{14}$$

3.3 CPR-BASED CARRIER CLASSIFICATION

After selecting the perturbation categories, we use the corresponding CPRs to classify carriers as real or fake. The pretrained visual foundation model remains frozen throughout this stage, and only the linear classification heads are trained using binary cross-entropy.

Single-Perturbation Classification. Given a carrier image $\mathbf{x}$ and a selected perturbation category ρ^*, we construct its CPR $\mathbf{r}_{\rho^*}(\mathbf{x})$ and feed it into a linear classification head:

$$p(\mathbf{x}) = \sigma\left(\text{Linear}\left(\mathbf{r}_{\rho^*}(\mathbf{x})\right)\right), \tag{15}$$

where $p(\mathbf{x})$ denotes the predicted probability that $\mathbf{x}$ is fake.

Multi-Perturbation Ensemble. The controllability of perturbations allows us to construct multiple perturbation branches. Specifically, we design several candidate perturbation-category groups from different perspectives:

$$\mathcal{G} = \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_M\}, \tag{16}$$

where each $\mathcal{R}_m$ is constructed according to a different perturbation perspective. Since different perturbation categories interact with the carrier in different ways, their CPRs may capture complementary response patterns.

The selection criterion is applied independently within each group to select one perturbation category:

$$\rho_m^* = \arg\max_{\rho \in \mathcal{R}_m} S(\rho), \qquad m = 1, \ldots, M. \tag{17}$$

For each selected category ρ_m^*, we construct the corresponding CPR and train an independent linear classification head:

$$p_m(\mathbf{x}) = \sigma\left(\text{Linear}_m\left(\mathbf{r}_{\rho_m^*}(\mathbf{x})\right)\right). \tag{18}$$

During inference, the outputs of these classification heads are averaged to obtain the ensemble probability:

$$p_{\text{ens}}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} p_m(\mathbf{x}), \qquad \widehat{y} = \mathbb{I}\left[p_{\text{ens}}(\mathbf{x}) \geq \tau\right], \tag{19}$$

where τ is the decision threshold. By aggregating predictions from different perturbation categories, the proposed method can further improve the performance of real–fake image classification.

Table 2: Quantitative evaluation of candidate perturbation categories. The selected perturbation category in each group is highlighted in bold.

Metric	$\mathcal{R}_{\mathrm{F}}$			$\mathcal{R}_{\mathrm{S}}$				$\mathcal{R}_{\mathrm{A}}$			
	$\rho_{\mathrm{F},1}$	$\rho_{\mathrm{F},2}$	$\boldsymbol{\rho_{\mathbf{F},3}}$	$\rho_{\mathrm{S},1}$	$\rho_{\mathrm{S},2}$	$\rho_{\mathrm{S},3}$	$\boldsymbol{\rho_{\mathbf{S},4}}$	$\rho_{\mathrm{A},1}$	$\rho_{\mathrm{A},2}$	$\rho_{\mathrm{A},3}$	$\boldsymbol{\rho_{\mathbf{A},4}}$
η_{car}	0.7283	0.5158	0.4459	0.3910	0.4042	0.3977	0.3799	0.5792	0.5114	0.4211	0.4026
η_{per}	0.0175	0.0805	0.1902	0.2203	0.1567	0.1765	0.0383	0.0829	0.0911	0.1035	0.0792
η_{int}	0.2542	0.4037	0.3638	0.3887	0.4391	0.4258	0.5818	0.3378	0.3975	0.4754	0.5181
$\widetilde{A}$	0.3513	0.5179	1.0000	1.0000	0.8942	0.9077	0.7795	1.0000	0.8956	0.7903	0.9291
$S(\rho)$	1.8248	2.1644	**2.3471**	2.3367	2.4590	2.4062	**2.8665**	2.5098	2.5085	2.5340	**2.8068**

Table 3: Performance comparison on the GenImage benchmark. All methods are trained on Stable Diffusion v1.4. The best and second-best results are highlighted in bold and underlined, respectively.

Method	ADM	BigGAN	Midjourney	VQDM	GLIDE	SD-v1.4	SD-v1.5	Wukong	Avg.
Gram-Net (Liu et al., 2020)	0.576	0.591	0.665	0.528	0.747	0.989	0.988	0.960	0.756
NPR (Tan et al., 2024)	0.619	0.849	0.850	0.567	0.844	0.996	0.994	0.984	0.838
UnivFD (Ojha et al., 2023)	0.571	0.839	0.795	0.644	0.842	0.959	0.958	0.926	0.817
SAFE (Li et al., 2025a)	0.600	0.880	0.829	0.708	0.836	0.964	0.967	0.903	0.836
Effort (Yan et al., 2024)	0.704	0.749	0.756	0.824	0.807	0.874	0.873	0.867	0.807
DDA (Chen et al., 2025)	0.880	0.741	**0.960**	0.719	0.862	0.986	0.985	0.986	0.890
OMAT (Zhou et al., 2025)	0.837	0.973	0.903	0.954	0.974	0.974	0.973	0.975	0.946
AIDE (Yan et al., 2025)	0.602	0.648	0.813	0.694	0.514	0.959	0.957	0.954	0.768
DINOv3-Linear (Zhou et al., 2026)	0.849	**0.991**	0.934	**0.992**	0.963	**0.998**	**0.996**	<u>0.994</u>	0.964
Ours									
CPR($\rho_{\mathrm{F},3}$)	**0.927**	0.983	<u>0.956</u>	0.987	<u>0.979</u>	0.989	0.987	0.987	<u>0.974</u>
CPR($\mathcal{R}^*$)	<u>0.916</u>	<u>0.989</u>	0.951	**0.992**	**0.987**	**0.998**	<u>0.995</u>	**0.995**	**0.978**

4 Experiments

4.1 Datasets

We conduct experiments on two categories of datasets: (1) **Standard Benchmarks:** We use GenImage (Zhu et al., 2023), a widely adopted benchmark comprising real images and images from eight generative models, including ADM (Dhariwal & Nichol, 2021), BigGAN (Brock et al., 2018), Midjourney, VQDM (Gu et al., 2022b), GLIDE (Nichol et al., 2021), Stable Diffusion v1.4, Stable Diffusion v1.5 (Rombach et al., 2022), and Wukong (Gu et al., 2022a). Following the official protocol (Ojha et al., 2023), we use the Stable Diffusion v1.4 for training. (2) **In-the-Wild Datasets:** We evaluate the generalization ability of our method on four datasets, namely Chameleon (Yan et al., 2025), WildRF (Cavia et al., 2024), SocialRF (Li et al., 2026), and SynthWildX (Cozzolino et al., 2024). These datasets contain diverse real and AI-generated images collected from online sources.

4.2 Perturbation Selection

Candidate Perturbation Groups. We design candidate perturbations from three perspectives: (1) the frequency-based perturbation group $\mathcal{R}_{\mathrm{F}}$, constructed by manipulating frequency-band components, including $\rho_{\mathrm{F},1}$, frequency-band noise injection; $\rho_{\mathrm{F},2}$, frequency-band carrier component addition; and $\rho_{\mathrm{F},3}$, frequency-band carrier filtering; (2) the local-structure perturbation group $\mathcal{R}_{\mathrm{S}}$, constructed by modifying local spatial relationships, including $\rho_{\mathrm{S},1}$, local structural inversion; $\rho_{\mathrm{S},2}$, local neighborhood permutation; $\rho_{\mathrm{S},3}$, edge continuity disruption; and $\rho_{\mathrm{S},4}$, patch wise noise injection; and (3) the encoder attention-guided perturbation group $\mathcal{R}_{\mathrm{A}}$, constructed according to the spatial importance indicated by the encoder, including $\rho_{\mathrm{A},1}$, attention-guided masked noise injection; $\rho_{\mathrm{A},2}$, attention-guided patch exchange; $\rho_{\mathrm{A},3}$, attention-guided semantic smoothing; and $\rho_{\mathrm{A},4}$, attention-guided masking.

Perturbation Evaluation. We evaluate the candidate perturbations using the ANOVA-inspired response decomposition method and perturbation selection criterion described in Sec. 3.2.1.

According to the overall selection scores, $\rho_{\mathrm{F},3}$, $\rho_{\mathrm{S},4}$, and $\rho_{\mathrm{A},4}$ are selected from the frequency-based, local-structure, and encoder attention-guided perturbation groups, respectively. The selected perturbation types are used in the subsequent CPR-based detector, with their specific rules provided

Table 4: Performance comparison on in-the-wild datasets. "Real" and "Fake" denote the accuracies on real and generated images, respectively, and "ACC" denotes the sample-proportion-weighted overall accuracy. "Avg." reports the mean across the four datasets. Bold and underlined values indicate the best and second-best results, respectively.

Method	Chameleon			WildRF			SocialRF			SynthWildX			Avg.		
	Real	Fake	ACC	Real	Fake	ACC	Real	Fake	ACC	Real	Fake	ACC	Real	Fake	ACC
Gram-Net	0.992	0.044	0.590	0.947	0.205	0.543	0.531	0.523	0.527	0.845	0.257	0.544	0.829	0.257	0.551
NPR	**0.999**	0.046	0.594	0.980	0.243	0.578	0.593	0.537	0.566	0.683	0.324	0.498	0.814	0.288	0.559
UnivFD	0.763	0.441	0.626	0.693	0.637	0.663	0.563	0.637	0.601	0.545	0.509	0.523	0.641	0.556	0.603
SAFE	0.993	0.046	0.590	0.918	0.309	0.586	0.564	0.532	0.549	0.772	0.283	0.521	0.812	0.293	0.562
Effort	0.394	0.782	0.559	0.179	0.955	0.602	0.513	0.533	0.523	0.345	0.889	0.616	0.358	0.790	0.575
DDA	0.940	0.708	0.841	0.899	0.908	0.904	0.818	0.846	0.832	0.936	0.894	0.909	0.898	0.839	0.872
OMAT	0.899	0.359	0.670	0.633	0.715	0.678	0.581	0.633	0.606	0.619	0.580	0.595	0.683	0.572	0.637
AIDE	0.944	0.203	0.629	0.973	0.195	0.549	0.578	0.541	0.561	0.690	0.297	0.488	0.796	0.309	0.557
DINOv3-Linear	0.933	**0.895**	0.917	0.948	0.975	0.963	0.937	0.948	0.942	0.917	0.913	0.909	0.934	0.933	0.933
Ours															
CPR($\rho_{\mathrm{F},3}$)	0.955	0.889	0.927	0.976	**0.976**	0.976	0.918	0.974	0.946	0.949	0.932	0.934	0.950	**0.943**	0.946
CPR($\mathcal{R}^*$)	0.977	0.869	**0.931**	**0.987**	**0.976**	**0.981**	**0.944**	**0.979**	**0.962**	**0.958**	**0.935**	**0.940**	**0.967**	0.940	**0.954**

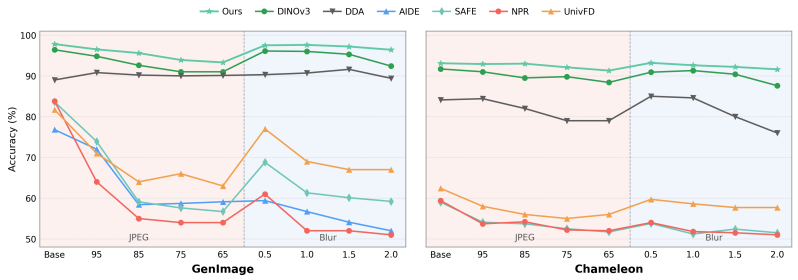


Figure 3: Robustness evaluation under JPEG compression and Gaussian blur on GenImage (left) and the in-the-wild Chameleon dataset (right).

in the appendix. The resulting perturbation set is

$$\mathcal{R}^* = \{\rho_{\mathrm{F},3}, \rho_{\mathrm{S},4}, \rho_{\mathrm{A},4}\}. \tag{20}$$

4.3 Quantitative Analysis

Tables 3 and 4 compare CPR with previous methods under generator and real-scene distribution shifts. We evaluate CPR with two inference schemes: CPR($\rho_{\mathrm{F},3}$), which uses the FFT-based single-perturbation classifier associated with $\rho_{\mathrm{F},3}$, and CPR($\mathcal{R}^*$), which employs multi-perturbation ensemble inference.

Performance on GenImage. CPR($\rho_{\mathrm{F},3}$) achieves an average accuracy of 97.4%, while CPR($\mathcal{R}^*$) reaches 97.8%. Compared with the strongest baseline DINOv3-Linear (96.4%), these results improve the average accuracy by 1.0% and 1.4%, respectively. This demonstrates that CPR maintains strong detection performance across unseen generation processes. The further improvement achieved by CPR($\mathcal{R}^*$) also suggests that combining multiple perturbation responses provides more effective and stable discriminative information.

Performance on in-the-wild datasets. The baseline DINOv3-Linear achieves an average real-image accuracy of 93.4%. CPR($\rho_{\mathrm{F},3}$) and CPR($\mathcal{R}^*$) improve this value to 95.0% and 96.7%, respectively, corresponding to gains of 1.7 and 3.2 percentage points. The average ACC increases from 93.3% for DINOv3-Linear to 94.6% and 95.4% for the two CPR variants, respectively. These results demonstrate that perturbation-induced responses improve cross-domain detection performance, providing complementary evidence for authenticity assessment.

4.4 Ablation Study and Robustness Evaluation

We conduct ablation studies under both the GenImage and in-the-wild settings. The experiments examine the effectiveness of the CPR representation, the perturbation selection strategy, and the joint use of multiple perturbations. The results are reported in Tables 5 and 6.

Table 5: Average accuracy of different perturbations and their joint combination on the GenImage and in-the-wild datasets. The perturbations $\rho_{F,3}$, $\rho_{S,4}$, and $\rho_{A,4}$ are the three perturbations selected in Sec. 4.2. $\mathcal{R}^*$ denotes the final result obtained by jointly combining these three perturbations.

Metric	$\mathcal{R}_F$			$\mathcal{R}_S$				$\mathcal{R}_A$				$\mathcal{R}^*$
	$\rho_{F,1}$	$\rho_{F,2}$	$\boldsymbol{\rho_{F,3}}$	$\rho_{S,1}$	$\rho_{S,2}$	$\rho_{S,3}$	$\boldsymbol{\rho_{S,4}}$	$\rho_{A,1}$	$\rho_{A,2}$	$\rho_{A,3}$	$\boldsymbol{\rho_{A,4}}$	
GenImage average acc.	0.798	0.924	**0.974**	0.854	0.946	0.884	**0.974**	0.926	0.888	0.945	**0.965**	0.978
In-the-wild average acc.	0.789	0.852	**0.946**	0.794	0.917	0.892	**0.941**	0.858	0.834	0.917	**0.938**	0.954

Table 6: Ablation study of different representations on the GenImage and in-the-wild datasets. The difference term $\Phi(x_i + \delta_i) - \Phi(x_i)$ is the final CPR representation adopted in this work. Each representation is evaluated using $\mathrm{acc}(\rho_{F,3})$ and $\mathrm{acc}(\mathcal{R}^*)$.

Dataset	$\Phi(x_i)$		$\Phi(x_i + \delta_i)$		$\Phi(\delta_i)$		$\boldsymbol{\Phi(x_i + \delta_i) - \Phi(x_i)}$	
	$\mathrm{acc}(\rho_{F,3})$	$\mathrm{acc}(\mathcal{R}^*)$	$\mathrm{acc}(\rho_{F,3})$	$\mathrm{acc}(\mathcal{R}^*)$	$\mathrm{acc}(\rho_{F,3})$	$\mathrm{acc}(\mathcal{R}^*)$	$\mathrm{acc}(\rho_{F,3})$	$\mathrm{acc}(\mathcal{R}^*)$
GenImage	0.964	0.964	0.901	0.916	0.885	0.894	0.974	**0.978**
In-the-wild	0.933	0.933	0.866	0.880	0.878	0.874	0.946	**0.954**

Effect of the CPR representation. Table 6 compares different representation choices. Directly using the perturbed-image representation $\Phi(x_i + \delta_i)$ or the perturbation representation $\Phi(\delta_i)$ leads to lower detection performance than using the original representation $\Phi(x_i)$. In contrast, the CPR representation $\Phi(x_i+\delta_i)-\Phi(x_i)$ achieves the best results. These results demonstrate that modeling CPR as a feature change is more effective than directly using the representations of the original image, the perturbed image, or the image-difference feature, namely, the feature representation of the perturbation itself.

Effect of perturbation selection. As shown in Table 5, the performance varies substantially across candidate perturbations. Within the three perturbation groups, $\rho_{\mathrm{F},3}$, $\rho_{\mathrm{S},4}$, and $\rho_{\mathrm{A},4}$ achieve the strongest or near-strongest performance, which is consistent with the categories selected by the criterion in Sec. 4.2. This consistency supports the effectiveness of the proposed perturbation selection strategy.The Appendix further analyzes the correlation between $S(\rho)$ and classification accuracy.

Effect of joint perturbation modeling. Combining the selected perturbations produces the best overall performance. The joint result $\mathcal{R}^*$ achieves 97.8% on GenImage and 95.4% in the in-the-wild setting, outperforming the corresponding single-perturbation results.This improvement indicates that combining the predictions derived from perturbations with different mechanisms enhances the robustness of CPR-based detection.

Robustness evaluation. We further evaluate detector robustness against two common post-processing operations, JPEG compression (quality $\in \{95, 85, 75, 65\}$) and Gaussian blur ($\sigma \in \{0.5, 1.0, 1.5, 2.0\}$), applied to the test images of GenImage and Chameleon. As shown in Fig. 3, our method remains the most stable across all settings, retaining 93.3% on GenImage at quality 65 and 91.4% on Chameleon, with a maximum drop of about 4.5 points from the unperturbed baseline.

5 Conclusion

In this work, we present Carrier-Perturbation Responses (CPR), an alternative basis for generalizable AI-generated image detection. CPR treats each image as a carrier and characterizes its response to a controlled perturbation through the feature change between the original and perturbed images. To identify effective perturbations, we develop an ANOVA-inspired selection criterion that favors perturbations with strong carrier-perturbation interactions, limited rule-induced effects, and sufficient response magnitude. Together, these designs shift the detection basis from scene-dependent appearance features extracted from the image itself toward perturbation-induced response patterns, providing complementary evidence for authenticity assessment across unseen scenes. Extensive experiments on standard benchmarks and in-the-wild datasets demonstrate the effectiveness and robustness of CPR in cross-domain detection.

AI USE STATEMENT

We declare that the use of large language models (LLMs) during the drafting of this manuscript was confined to language-related assistance, such as sentence refinement and grammatical corrections. All substantive content was independently authored by the authors and underwent rigorous review and verification following any modifications based on LLM assistance. This research did not involve any other processes reliant upon large language models.

REFERENCES

Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–7. IEEE, 2018.

Lorenzo Baraldi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, Alessandro Nicolosi, and Rita Cucchiara. Contrasting deepfakes diffusion via contrastive learning and global-local similarities. In *European Conference on Computer Vision*, pp. 199–216. Springer, 2024.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, et al. Perception encoder: The best visual embeddings are not at the output of the network. *Advances in Neural Information Processing Systems*, 38:60884–60937, 2026.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world. *arXiv preprint arXiv:2406.09398*, 2024.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, et al. Dual data alignment makes ai-generated image detector easier generalizable. *arXiv preprint arXiv:2505.14359*, 2025.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, Jim Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *Advances in Neural Information Processing Systems*, 38:48009–48036, 2026.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4356–4366. IEEE, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Ronald Aylmer Fisher. *Statistical methods for research workers*. Number 5. Oliver and Boyd, 1930.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pp. 3247–3258. PMLR, 2020.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431, 2022a.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10686–10696. IEEE, 2022b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4396–4405. IEEE, 2019.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining v. 1*, pp. 2405–2414, 2025a.

Yiheng Li, Yang Yang, Zichang Tan, Huan Liu, Weihua Chen, Xu Zhou, and Zhen Lei. Unleashing the potential of consistency learning for detecting and grounding multi-modal media manipulation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9243–9252. IEEE, 2025b.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? *Advances in neural information processing systems*, 38, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10770–10780. IEEE, 2024.

Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8057–8066. IEEE, 2020.

Douglas C Montgomery. *Design and analysis of experiments, Minitab manual*. John wiley & sons, 2010.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489. IEEE, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9130–9140. IEEE, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28130–28139, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8692–8701. IEEE, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22388–22398. IEEE, 2023.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*, volume 2025, pp. 70702–70720, 2025.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024.

Y Zhou, X He, K Lin, B Fan, F Ding, B Li, et al. Breaking latent prior bias in detectors for generalizable aigc image detection. arxiv. *arXiv preprint arXiv:2506.00874*, 2025.

Yue Zhou, Xinan He, Kaiqing Lin, Bing Fan, Feng Ding, and Bin Li. Simplicity prevails: The emergence of generalizable aigi detection in visual foundation models. *arXiv preprint arXiv:2602.01738*, 2026.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in neural information processing systems*, 36:77771–77782, 2023.

A Implementation Details

Our detector uses DINOv3-7B as the visual feature extractor. During training, all parameters of DINOv3-7B are frozen, and only a single linear classification head is optimized for binary real-fake image classification. For perturbation selection, we use the Stable Diffusion v1.4 training split, evaluate real and fake images separately, and equally average their variance-based statistics. The main experiments use $(w_{\mathrm{int}}, w_{\mathrm{car}}, w_{\mathrm{per}}, \lambda) = (3, 1, -1, 1)$, while the sensitivity to these weighting coefficients is evaluated in the appendix. The model is trained with a learning rate of 0.002, a batch size of 128, and 6 training epochs, and $\tau = 0.5$ is used during inference. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU.

To improve training efficiency, we decouple CPR construction from classifier optimization. In the first stage, we use the frozen DINOv3-7B to construct the CPR for each training image under the corresponding perturbation and store the resulting CPRs together with their binary labels. In the second stage, we load the cached CPR-label pairs and optimize the linear classification head using the cross-entropy loss. Since the CPRs are independent of the classifier parameters, this two-stage procedure avoids repeatedly reconstructing CPRs during training and reduces the subsequent optimization to lightweight linear classification. With this procedure, the complete training pipeline can be finished within approximately eight hours on a single NVIDIA GeForce RTX 4090 GPU.

B Details of the Datasets

To comprehensively evaluate the generalization ability of CPR-based detection, we conduct experiments on 13 evaluation subsets, including eight generator-specific test subsets from GenImage and four in-the-wild datasets. In addition, the Stable Diffusion v1.4 training split is used for detector training and perturbation selection. These datasets cover different image sources, generative models, collection platforms, and post-processing conditions. Detailed dataset statistics are summarized in Table 7.

ImageNet. ImageNet is a large-scale image database organized according to the WordNet hierarchy. Its images are collected from the web and categorized into a wide range of object synsets, providing diverse natural scenes, objects, and visual appearances. In GenImage, ImageNet images serve as the real-image source.

Microsoft COCO. Microsoft COCO is constructed from real-world photographs collected from Flickr and focuses on common objects and everyday scenes. The dataset provides dense annotations, including object instances, captions, and human keypoints, which result in diverse compositions and object-centric visual content.

Open Images. Open Images is a large-scale real-image collection gathered from Flickr. It covers diverse scenes and object categories and is accompanied by image-level labels, object bounding boxes, and visual-relation annotations. Compared with ImageNet and COCO, it contains a broader range of unconstrained image content.

GenImage. GenImage is a million-scale benchmark constructed by combining real images from ImageNet with synthetic images generated by eight representative generative models. Specifically, it contains generator-specific subsets based on ADM, BigGAN, Midjourney, Vector Quantized Diffusion Model (VQDM), GLIDE, Stable Diffusion v1.4, Stable Diffusion v1.5, and Wukong. For each subset, real images are drawn from ImageNet, while fake images are produced by the corresponding generator under semantically related image categories. This construction provides a common real-image source and multiple synthetic distributions, enabling systematic evaluation of cross-generator generalization.

Stable Diffusion v1.4 Training Split. The Stable Diffusion v1.4 training split is the training partition of the corresponding GenImage subset. It consists of real images from ImageNet and synthetic images generated by Stable Diffusion v1.4. The two classes share the same broad semantic source while differing in image-generation process, making this split suitable for perturbation selection and classifier training.

Chameleon. Chameleon is an in-the-wild benchmark constructed to reflect realistic image-generation and image-sharing conditions. Its fake images are collected from AI-art communities

Table 7: Statistics of the datasets used in our experiments.

Dataset	Real / Fake	Source
GenImage-ADM (Test)	6K / 6K	ImageNet + ADM
GenImage-BigGAN (Test)	6K / 6K	ImageNet + BigGAN
GenImage-Midjourney (Test)	6K / 6K	ImageNet + Midjourney
GenImage-VQDM (Test)	6K / 6K	ImageNet + VQDM
GenImage-GLIDE (Test)	6K / 6K	ImageNet + GLIDE
GenImage-Stable Diffusion v1.4 (Test)	6K / 6K	ImageNet + Stable Diffusion v1.4
GenImage-Stable Diffusion v1.5 (Test)	8K / 8K	ImageNet + Stable Diffusion v1.5
GenImage-Wukong (Test)	6K / 6K	ImageNet + Wukong
GenImage-Stable Diffusion v1.4 (Train)	162K / 162K	ImageNet + Stable Diffusion v1.4
Chameleon	14.9K / 11.2K	Public web + ArtStation + Civitai + Liblib
WildRF	2650 / 2650	Reddit + Facebook + X
SocialRF	3K / 3K	X + Facebook + Reddit
SynthWildX	500 / 1.5K	X

and platforms, including ArtStation, Civitai, and Liblib, and include images generated by systems such as Midjourney, DALL-E 3, and Stable Diffusion-based models with customized LoRA configurations. The real images are collected from public web pages using search queries related to the synthetic-image content, which helps reduce semantic differences between the real and fake classes. The resulting images are further curated to retain visually convincing samples that are difficult to distinguish through human inspection.

WildRF. WildRF is constructed from public image posts on Reddit, Facebook, and X. Instead of generating all fake images under a single controlled protocol, the dataset collects images that naturally appear in social-media environments, including authentic photographs and AI-generated images shared by users. Consequently, the images exhibit platform-dependent resizing, compression, reposting, and other uncontrolled variations. This construction makes WildRF representative of real-world social-media image distributions.

SocialRF. SocialRF is collected by crawling public posts and images from X, Facebook, and Reddit. Its real and synthetic samples are identified through platform content and targeted collection rules, using tags such as #nature, #photography, and #realphoto for authentic images and tags such as #aiart, #aigenerated, and #fakephoto for synthetic images. The resulting data contain naturally occurring variations in image quality, compression, and presentation style, reflecting the distribution shifts encountered in social-media content.

SynthWildX. SynthWildX is curated from image content shared on X. It contains authentic images and synthetic images posted in the wild, rather than images generated and collected under a unified laboratory setting. The synthetic portion consequently includes images produced by different contemporary generation systems, while both classes inherit the resizing, compression, reposting, and stylistic variations of the X platform.

C Detailed Implementation of Candidate Perturbations

This section details the implementation of the three perturbation categories selected for the final detector, namely $\rho_{F,3}$, $\rho_{S,4}$, and $\rho_{A,4}$. For each category, we first describe the corresponding perturbation rule and its configurable parameters, and then specify the parameter values used in the final classification experiments.

Following the notation introduced in the main text, we denote a perturbation category by ρ and its corresponding rule parameters by ξ.

C.1 Frequency-Band Carrier Filtering.

The perturbation $\rho_{F,3}$ modifies the frequency composition of an image by retaining a selected band of Fourier components and suppressing the remaining components. Given an input image x_i of height H and width W, we first compute its two-dimensional Fourier transform and shift the zero-

frequency component to the center of the spectrum. For a frequency coordinate (u, v), let (u_c, v_c) denote the center of the shifted spectrum. We define the normalized radial frequency as

$$d(u,v) = \frac{\sqrt{(u-u_c)^2 + (v-v_c)^2}}{\max\limits_{u',v'} \sqrt{(u'-u_c)^2 + (v'-v_c)^2}}, \tag{21}$$

where $d(u, v) \in [0, 1]$. A frequency band is specified by a lower boundary ℓ_F and an upper boundary u_F, with

$$0 \le \ell_F < u_F \le 1. \tag{22}$$

The corresponding band-pass mask is defined as

$$M_{F,3}(u,v;\ell_F,u_F) = \begin{cases} 1, & \ell_F \le d(u,v) \le u_F, \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

The filtered image is reconstructed as

$$\tilde{x}_i^{F,3} = \mathcal{F}^{-1}\left(M_{F,3} \odot \mathcal{F}(x_i)\right), \tag{24}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier and inverse Fourier transforms, respectively.

The rule parameters are represented as

$$\xi_{F,3} = (\ell_F, b_F), \tag{25}$$

where ℓ_F is the lower boundary of the retained frequency band and b_F is its bandwidth. The upper boundary is determined by $u_F = \ell_F + b_F$. During candidate-rule construction, both the location and the bandwidth of the retained frequency band are randomly sampled subject to the valid frequency range. In the final classification experiments, we use

$$\ell_F = 0.7, \qquad u_F = 1.0, \tag{26}$$

which corresponds to retaining the normalized high-frequency band $[0.7, 1.0]$ and suppressing frequency components outside this band.

C.2 Patch-Wise Noise Injection.

The perturbation $\rho_{S,4}$ introduces localized stochastic changes by replacing a subset of non-overlapping image patches with Gaussian noise. Given an image of height H and width W, the image is divided into square patches with side length s_S. Let P denote the total number of patches, and let α_S denote the proportion of patches to be replaced. The number of selected patches is

$$N_S = \lfloor \alpha_S P \rfloor. \tag{27}$$

The N_S patches are sampled uniformly at random. For each selected patch, the original pixel values are replaced with Gaussian noise sampled according to

$$\epsilon_i \sim \mathcal{N}\left(\mu_{x_i}, \sigma_{x_i}^2\right), \tag{28}$$

where μ_{x_i} and σ_{x_i} denote the mean and standard deviation of the input image tensor, respectively. The unselected patches remain unchanged. If $\mathcal{S}_i$ denotes the set of selected patches, the perturbed image can be written as

$$\tilde{x}_i^{S,4}(p) = \begin{cases} \epsilon_i(p), & p \in \mathcal{S}_i, \\ x_i(p), & p \notin \mathcal{S}_i. \end{cases} \tag{29}$$

The rule parameters are

$$\xi_{S,4} = (s_S, \alpha_S), \tag{30}$$

where s_S controls the spatial granularity of the perturbation and α_S controls the proportion of corrupted patches. In the final classification experiments, we set

$$s_S = 16, \qquad \alpha_S = 0.3. \tag{31}$$

Thus, the perturbation replaces 30% of the image patches with Gaussian noise while preserving the remaining patches.

C.3 ATTENTION-GUIDED MASKING.

The perturbation $\rho_{A,4}$ masks spatial regions according to the attention distribution of the frozen visual encoder. We use the same patch grid as DINOv3-7B in the original-image branch, with each patch covering 16×16 pixels in the encoder input.

For the original image x_i, we extract the attention weights from the last encoder layer and average the CLS-to-spatial-patch attention scores over all attention heads. Only spatial patch tokens are included in the ranking; the CLS token and register tokens are excluded.

Let P denote the number of spatial patch tokens and let α_A denote the masking proportion. The number of patches to mask is

$$N_A = \lfloor \alpha_A P \rfloor . \tag{32}$$

Let $a_i(p)$ denote the head-averaged attention score from the CLS token to spatial patch token p. We define the selected index set as

$$\mathcal{A}_i = \mathrm{TopKIndices}\left(\{a_i(p)\}_{p=1}^{P}, N_A\right), \tag{33}$$

where $\mathrm{TopKIndices}$ returns the indices of the N_A largest scores.

We set all pixel values in the corresponding selected regions to zero and leave the remaining regions unchanged:

$$\tilde{x}_i^{A,4}(p) = \begin{cases} \mathbf{0}, & p \in \mathcal{A}_i, \\ x_i(p), & p \notin \mathcal{A}_i, \end{cases} \tag{34}$$

where $x_i(p)$ denotes the pixel block corresponding to spatial patch token p in the encoder input tensor.

The perturbation rule is parameterized by the masking proportion:

$$\xi_{A,4} = \alpha_A. \tag{35}$$

In the final classification experiments, we set $\alpha_A = 0.3$, masking approximately 30% of the spatial patches with the highest CLS attention scores while preserving the remaining image content.

C.4 PARAMETER SELECTION AND RANDOMNESS.

The parameters used in the final classification experiments were manually specified as fixed configurations for instantiating the three selected perturbations. We investigate whether the experimental performance depends strongly on these manually specified parameters. As shown in Table 8, the main-experiment configurations remain competitive within the evaluated parameter ranges.

When the input image, encoder parameters, and perturbation parameters are fixed, Frequency-Band Carrier Filtering and Attention-Guided Masking are deterministic. In contrast, Patch-Wise Noise Injection is stochastic: the patch locations to be replaced are sampled uniformly from all image patches, and the replacement pixel values are independently sampled from a Gaussian distribution whose mean and standard deviation are computed from the input image. We fixed the training random seed to 42. For each parameter configuration of Patch-Wise Noise Injection, we evaluated the test set using ten perturbation seeds, $\{10, 20, \ldots, 100\}$, and report the mean accuracy and standard deviation across these runs. The reported standard deviation therefore reflects the performance variation introduced by stochastic noise perturbations, rather than changes in the encoder or perturbation category.

D ANALYSIS OF WEIGHTING COEFFICIENTS

This section examines whether the priority relation $w_{\mathrm{int}} > w_{\mathrm{car}} > w_{\mathrm{per}}$ introduced in Sec. 3.2.2 is consistent with downstream detection performance. We first compare different assignments of the main weight values and then vary their spacings to assess whether the preferred ordering remains effective beyond the main configuration.

Table 8: Sensitivity to perturbation parameters. All reported values are classification accuracies (Acc) evaluated on the Chameleon dataset. Panels (a) and (b) report one evaluation per configuration, whereas panel (c) reports mean ± standard deviation over repeated evaluations with different perturbation seeds. In panel (a), for bandwidth b, High, Mid, and Low denote $[1-b, 1]$, $[0.5-b/2, 0.5+b/2]$, and $[0, b]$, respectively. † denotes the configuration used in the main experiments.

(a) Frequency-Band Carrier Filtering

Bandwidth	High	Mid	Low
0.15	0.929	0.925	0.928
0.30	$0.927^{\dagger}$	0.921	0.925
0.45	0.924	0.915	0.927
0.60	0.915	0.913	0.916
0.75	0.912	0.914	0.910

(b) Attention-Guided Masking

Masking proportion	Accuracy
0.2	0.917
0.3	$0.915^{\dagger}$
0.4	0.915

(c) Patch-Wise Noise Injection

	Replacement proportion		
Patch size	0.2	0.3	0.4
16	0.916990 ± 0.000423	$(0.917813 \pm 0.000351)^{\dagger}$	0.917182 ± 0.000513
14	0.915876 ± 0.000374	0.918028 ± 0.000289	0.916656 ± 0.000446

D.1 Analysis Setup and Metrics

Setup. We use the category statistics in Table 2 and the average classification accuracies in Table 5. The analysis includes all eleven individual categories in $\mathcal{R}_{\mathrm{F}}$, $\mathcal{R}_{\mathrm{S}}$, and $\mathcal{R}_{\mathrm{A}}$, containing three, four, and four candidates, respectively.

For each weight configuration, we recompute $S(\rho_k)$ using Eq. equation 13. The effect contributions $\eta_q^{(k)}$, normalized CPR strength $\widetilde{A}^{(k)}$, and measured accuracies remain fixed. We set $\lambda = 1$ throughout and vary only w_{int}, w_{car}, and w_{per}. Thus, this analysis does not require retraining the individual detectors.

Evaluation metrics. Following Eq. equation 17, we evaluate rankings independently within each group rather than pooling raw scores across groups. Let $\mathrm{acc}_d(\rho)$ denote the average accuracy of category ρ under evaluation setting d, corresponding to GenImage or in-the-wild evaluation. For each weight configuration, we compute

$$\begin{aligned} r_{s,m}^{(d)} &= \mathrm{Spearman}\left(\big(S(\rho)\big)_{\rho\in\mathcal{R}_m}, \big(\mathrm{acc}_d(\rho)\big)_{\rho\in\mathcal{R}_m}\right), \\ \bar{r}_s^{(d)} &= \frac{1}{M}\sum_{m=1}^{M} r_{s,m}^{(d)}, \qquad M = 3. \end{aligned} \tag{36}$$

The score and accuracy vectors use the same category order, and the three groups receive equal weight in the mean correlation.

Table 9 reports the accuracy rankings. Although the accuracy values differ between the two evaluation settings, their within-group rankings are identical. Their Spearman correlations with any given score ranking are therefore also identical. For brevity, the tables omit the evaluation-setting superscript and use r_s for a within-group correlation and $\bar{r}_s$ for its mean over the three groups. These values apply to both settings. The two accuracy vectors are not averaged before computing correlations.

We additionally report a Top-1 selection count: the number of groups in which the category selected by Eq. equation 17 also achieves the highest measured accuracy among the evaluated candidates. The count is expressed as a fraction of the three groups. It is identical for the two evaluation settings because their highest-accuracy categories coincide. This metric evaluates individual-category selection, not the performance of a newly formed ensemble.

Table 9: Within-group accuracy rankings. Each entry lists categories in descending order of average accuracy for the indicated evaluation setting. The symbol $>$ denotes higher accuracy.

Accuracy metric	$\mathcal{R}_{\mathrm{F}}$	$\mathcal{R}_{\mathrm{S}}$	$\mathcal{R}_{\mathrm{A}}$
GenImage acc.	$\rho_{\mathrm{F},3} > \rho_{\mathrm{F},2} > \rho_{\mathrm{F},1}$	$\rho_{\mathrm{S},4} > \rho_{\mathrm{S},2} > \rho_{\mathrm{S},3} > \rho_{\mathrm{S},1}$	$\rho_{\mathrm{A},4} > \rho_{\mathrm{A},3} > \rho_{\mathrm{A},1} > \rho_{\mathrm{A},2}$
In-the-wild acc.	$\rho_{\mathrm{F},3} > \rho_{\mathrm{F},2} > \rho_{\mathrm{F},1}$	$\rho_{\mathrm{S},4} > \rho_{\mathrm{S},2} > \rho_{\mathrm{S},3} > \rho_{\mathrm{S},1}$	$\rho_{\mathrm{A},4} > \rho_{\mathrm{A},3} > \rho_{\mathrm{A},1} > \rho_{\mathrm{A},2}$

Table 10: Comparison of all six permutations of $\{3, 1, -1\}$, with $\lambda = 1$. Weights are ordered as $(w_{\mathrm{int}}, w_{\mathrm{car}}, w_{\mathrm{per}})$. Panel (a) gives descending score rankings: $\rho_i > \rho_j$ means $S(\rho_i) > S(\rho_j)$. Panel (b) gives the within-group Spearman correlations r_s between scores and accuracy. $\bar{r}_s$ is their equal-weight mean over the three groups. Correlations apply to both evaluation settings.

(a) Score rankings

Weights	$\mathcal{R}_{\mathrm{F}}$	$\mathcal{R}_{\mathrm{S}}$	$\mathcal{R}_{\mathrm{A}}$
$\mathbf{(3, 1, -1)}$	$\rho_{\mathrm{F},3} > \rho_{\mathrm{F},2} > \rho_{\mathrm{F},1}$	$\rho_{\mathrm{S},4} > \rho_{\mathrm{S},2} > \rho_{\mathrm{S},3} > \rho_{\mathrm{S},1}$	$\rho_{\mathrm{A},4} > \rho_{\mathrm{A},3} > \rho_{\mathrm{A},1} > \rho_{\mathrm{A},2}$
$(3, -1, 1)$	$\rho_{\mathrm{F},3} > \rho_{\mathrm{F},2} > \rho_{\mathrm{F},1}$	$\rho_{\mathrm{S},4} > \rho_{\mathrm{S},1} > \rho_{\mathrm{S},2} > \rho_{\mathrm{S},3}$	$\rho_{\mathrm{A},4} > \rho_{\mathrm{A},3} > \rho_{\mathrm{A},2} > \rho_{\mathrm{A},1}$
$(1, 3, -1)$	$\rho_{\mathrm{F},1} > \rho_{\mathrm{F},3} > \rho_{\mathrm{F},2}$	$\rho_{\mathrm{S},4} > \rho_{\mathrm{S},2} > \rho_{\mathrm{S},3} > \rho_{\mathrm{S},1}$	$\rho_{\mathrm{A},1} > \rho_{\mathrm{A},2} > \rho_{\mathrm{A},4} > \rho_{\mathrm{A},3}$
$(1, -1, 3)$	$\rho_{\mathrm{F},3} > \rho_{\mathrm{F},2} > \rho_{\mathrm{F},1}$	$\rho_{\mathrm{S},1} > \rho_{\mathrm{S},3} > \rho_{\mathrm{S},2} > \rho_{\mathrm{S},4}$	$\rho_{\mathrm{A},4} > \rho_{\mathrm{A},3} > \rho_{\mathrm{A},2} > \rho_{\mathrm{A},1}$
$(-1, 3, 1)$	$\rho_{\mathrm{F},1} > \rho_{\mathrm{F},3} > \rho_{\mathrm{F},2}$	$\rho_{\mathrm{S},1} > \rho_{\mathrm{S},3} > \rho_{\mathrm{S},2} > \rho_{\mathrm{S},4}$	$\rho_{\mathrm{A},1} > \rho_{\mathrm{A},2} > \rho_{\mathrm{A},4} > \rho_{\mathrm{A},3}$
$(-1, 1, 3)$	$\rho_{\mathrm{F},3} > \rho_{\mathrm{F},1} > \rho_{\mathrm{F},2}$	$\rho_{\mathrm{S},1} > \rho_{\mathrm{S},3} > \rho_{\mathrm{S},2} > \rho_{\mathrm{S},4}$	$\rho_{\mathrm{A},1} > \rho_{\mathrm{A},2} > \rho_{\mathrm{A},4} > \rho_{\mathrm{A},3}$

(b) Rank correlations

Weights	Within-group Spearman r_s			$\bar{r}_s$
	$\mathcal{R}_{\mathrm{F}}$	$\mathcal{R}_{\mathrm{S}}$	$\mathcal{R}_{\mathrm{A}}$	
$\mathbf{(3, 1, -1)}$	**1.000**	**1.000**	**1.000**	**1.000**
$(3, -1, 1)$	1.000	0.400	0.800	0.733
$(1, 3, -1)$	-0.500	1.000	-0.600	-0.033
$(1, -1, 3)$	1.000	-1.000	0.800	0.267
$(-1, 3, 1)$	-0.500	-1.000	-0.600	-0.700
$(-1, 1, 3)$	0.500	-1.000	-0.600	-0.367

D.2 Comparison of Weight Priority Orderings

We enumerate all six assignments of the main weight values $\{3, 1, -1\}$ to the interaction, carrier, and perturbation effects. The available weight values and λ remain fixed. Only the assignment of weights to effects changes. Throughout this section, weight tuples are ordered as $(w_{\mathrm{int}}, w_{\mathrm{car}}, w_{\mathrm{per}})$, so each tuple uniquely specifies the priority relation.

Table 10 presents the comparison in two panels. Panel (a) shows the descending score rankings using the original perturbation notation. Panel (b) reports the corresponding Spearman correlations with the accuracy rankings in Table 9.

With the proposed assignment $(3, 1, -1)$, the score rankings exactly match the accuracy rankings in all three groups, yielding $r_s = 1.000$ in each group. The criterion therefore recovers the complete within-group ordering rather than only identifying the highest-accuracy category.

Keeping the interaction effect at the highest priority but reversing the other two priorities, as in $(3, -1, 1)$, reduces $\bar{r}_s$ to 0.733. This alternative still selects the highest-accuracy category in every group but misorders some of the remaining candidates. The other four assignments yield $\bar{r}_s$ values between -0.700 and 0.267. Thus, the proposed priority relation achieves the strongest overall ranking agreement under the main weight values.

Table 11: Sensitivity to weight spacing with $\lambda = 1$. Weights are listed as $(w_{\text{int}}, w_{\text{car}}, w_{\text{per}})$ and follow the proposed priority relation. $\bar{r}_s$ denotes the mean within-group Spearman correlation, as in Table 10(b). Best? indicates whether this assignment has the uniquely highest $\bar{r}_s$ among all six permutations at the same gap configuration. Top-1 counts the groups whose selected category achieves the highest observed accuracy, out of three. Results apply to both evaluation settings.

d_1	d_2	Weights	$\bar{r}_s$	Best?	Top-1
0.5	0.5	$(1, 0.5, 0)$	0.400	Yes	2/3
0.5	1	$(1.5, 1, 0)$	0.600	Yes	3/3
0.5	2	$(2.5, 2, 0)$	0.800	Yes	3/3
1	0.5	$(1.5, 0.5, 0)$	0.600	Yes	3/3
1	1	$(2, 1, 0)$	0.733	Yes	3/3
1	2	$(3, 2, 0)$	0.800	Yes	3/3
2	0.5	$(2.5, 0.5, 0)$	0.933	Yes	3/3
2	1	$(3, 1, 0)$	0.933	Yes	3/3
2	2	$(4, 2, 0)$	1.000	Yes	3/3

D.3 Sensitivity to Weight Spacing

Spacing configurations. We next test whether the preferred priority relation depends on the particular values $\{3, 1, -1\}$. Let $w_{\text{H}} > w_{\text{M}} > w_{\text{L}}$ denote the high, middle, and low weight values. Their adjacent gaps are

$$d_1 = w_{\text{H}} - w_{\text{M}}, \qquad d_2 = w_{\text{M}} - w_{\text{L}}, \qquad d_1, d_2 \in \{0.5, 1, 2\}. \tag{37}$$

For each of the nine gap configurations, we use

$$(w_{\text{H}}, w_{\text{M}}, w_{\text{L}}) = (d_1 + d_2, d_2, 0) \tag{38}$$

and evaluate all six assignments to the three effects. This gives 54 configurations with $\lambda = 1$. The grid covers both equal and unequal adjacent gaps and varies the strength of the effect-weighted term relative to the amplitude term.

Setting the minimum weight to zero avoids repeatedly testing common-offset variants. Under the normalization in Sec. 3.2.1, the three effect contributions sum to approximately one. A common shift of the three weights therefore approximately adds the same constant to all scores. This equivalence is exact without the numerical stabilizer and approximate when its contribution is negligible. The representative weights $(4, 2, 0)$ have the same gaps as the main weights $(3, 1, -1)$ and produce the same rankings in the reported calculations.

Results. Table 11 summarizes the proposed priority relation at each grid point, assigning

$$(w_{\text{int}}, w_{\text{car}}, w_{\text{per}}) = (d_1 + d_2, d_2, 0). \tag{39}$$

For every row, its $\bar{r}_s$ is compared with those of the other four assignments of the same weight values.

The proposed priority relation achieves the uniquely highest $\bar{r}_s$ among the six assignments at all nine grid points, with values ranging from 0.400 to 1.000. In eight configurations, it also selects $\rho_{\text{F},3}$, $\rho_{\text{S},4}$, and $\rho_{\text{A},4}$, the highest-accuracy candidates in both evaluation settings.

The exception is the smallest-gap configuration, $(d_1, d_2) = (0.5, 0.5)$, represented by $(1, 0.5, 0)$. It has the same gaps as $(2, 1.5, 1)$ and selects $\rho_{\text{S},1}$ instead of $\rho_{\text{S},4}$, while retaining the best candidates in the other two groups.

E Metrics for Cross-Domain Consistency and Real–Fake Separability

We quantify cross-domain consistency and real–fake separation using five metrics. Each evaluation uses 500 images from each of four groups: seen real, seen fake, unseen real, and unseen fake, yielding 2,000 images in total. Original features and CPRs are evaluated on the same images within each repetition. All computations are performed in the 8192-dimensional feature spaces before t-SNE, with each feature vector independently L2-normalized. We repeat the evaluation 20 times, compute all five metrics separately in each repetition, and report their means and standard deviations.

Notation and centroids. Let $\mathcal{Z}_y^d$ denote the normalized features of class $y \in \{r, f\}$ in domain $d \in \{s, u\}$, where r/f indicate real/fake and s/u indicate seen/unseen. For a feature vector h, its normalized representation is $z = h/\|h\|_2$. The class-specific domain centroids are

$$c_y^d = \frac{1}{|\mathcal{Z}_y^d|} \sum_{z \in \mathcal{Z}_y^d} z. \tag{40}$$

For each class, we pool seen and unseen samples to obtain

$$c_y = \frac{\sum_{z \in \mathcal{Z}_y^s} z + \sum_{z \in \mathcal{Z}_y^u} z}{|\mathcal{Z}_y^s| + |\mathcal{Z}_y^u|}. \tag{41}$$

Since the two domains contain equal numbers of samples, $c_y = (c_y^s + c_y^u)/2$ in our evaluation. Centroids are not L2-normalized again after averaging.

Cross-domain centroid distances. We measure the displacement between seen and unseen centroids separately for each class:

$$D_{\text{real}}^{\text{cross}} = \|c_r^s - c_r^u\|_2, \qquad D_{\text{fake}}^{\text{cross}} = \|c_f^s - c_f^u\|_2. \tag{42}$$

Smaller values indicate less cross-domain displacement of the corresponding class centroid. These metrics measure mean shifts, rather than full distributional agreement or within-class compactness.

Real–fake centroid distance. We measure the distance between the pooled real and fake centroids:

$$D_{\text{class}} = \|c_r - c_f\|_2. \tag{43}$$

A larger value indicates greater real–fake centroid separation, but does not alone establish sample-level separability.

Shift-specific AUROCs. To evaluate cross-domain discrimination, we use a nearest-centroid score based only on seen-domain centroids:

$$s(z) = \|z - c_r^s\|_2 - \|z - c_f^s\|_2. \tag{44}$$

Higher scores indicate stronger evidence for the fake class. This scorer requires no gradient-based optimization, and unseen samples do not contribute to its centroids. The pooled centroids c_r and c_f used above are not used for scoring.

We compute two AUROCs by changing which class is drawn from the unseen domain:

$$\text{AUROC}_{\text{real}} = \text{AUC}\big(\{(s(z), 0) : z \in \mathcal{Z}_r^u\} \cup \{(s(z), 1) : z \in \mathcal{Z}_f^s\}\big), \tag{45}$$

$$\text{AUROC}_{\text{fake}} = \text{AUC}\big(\{(s(z), 0) : z \in \mathcal{Z}_r^s\} \cup \{(s(z), 1) : z \in \mathcal{Z}_f^u\}\big). \tag{46}$$

Real and fake samples have labels 0 and 1, respectively. The subscripts identify the class undergoing domain shift, not the positive class: fake is the positive class in both evaluations. Within each repetition, the same scoring function is used for both evaluation sets.

F Discussion

Our experiments show that CPR generalizes better to real scenes unseen during training. One possible explanation is that images from different real-scenes, despite differing in content and appearance, share low-level structures shaped by the real image formation process and natural image statistics. Applying the same category and rule controlled perturbations to different images allows CPR to capture the interaction between image representations and perturbations. In this process, the pretrained visual foundation model provides a representation space shaped by statistical regularities learned from large-scale natural images, which are broadly shared across real images. Scene changes may affect individual responses without substantially changing their overall distribution.

Generated images may resemble real images in semantics and appearance, but their local structures, texture organization, or frequency statistics can deviate from those of real images. They therefore produce different responses to the same perturbations category. CPR may consequently learn interaction patterns that are less dependent on scene content and remain informative for distinguishing real from generated images across unseen scenes.

G LIMITATIONS

Our study evaluates the proposed weighting strategy within a predefined library of candidate perturbations. Although the library covers several types of perturbations, including frequency-domain, spatial, stochastic, and attention-guided perturbations, other possible perturbation mechanisms remain to be explored.

Our analysis focuses on selecting perturbation categories within each group and does not further investigate the joint use or adaptive combination of multiple perturbations. In the current framework, CPRs from multiple perturbations are integrated through a direct aggregation scheme. Future work could explore more effective fusion strategies for multi-perturbation CPRs, such as methods that better capture the complementary information and interactions among different perturbations.